Speech-to-Speech Translation

Dr. Yuqing Gao
Research Staff Member & Manager
Speech Recognition, Understanding and Translation

IBM T. J. Watson Research Center

Thursday, July 22, 2004

An Overview of IBM Research

The Sun Never Sets at IBM Research

Almaden

Established: 1986
Employees: 500

Watson

Established: 1961
Employees: 1750

Zürich

Established: 1955
Employees: 300

Beijing

Established: 1995
Employees: 90

Austin

Established: 1995
Employees: 40

Haifa

Established: 1972
Employees: 500

Delhi

Established: 1998
Employees: 60

Tokyo

Established: 1982
Employees: 200

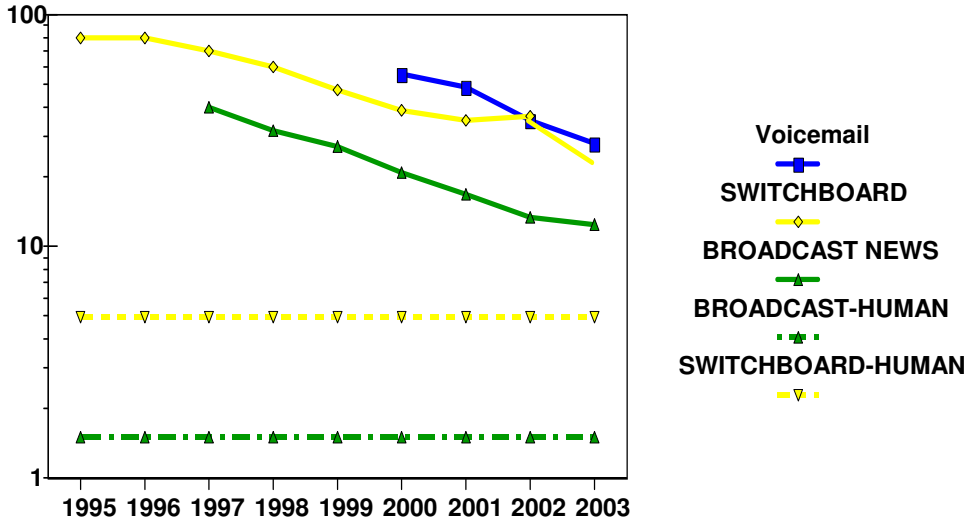
Major Conversational Technologies

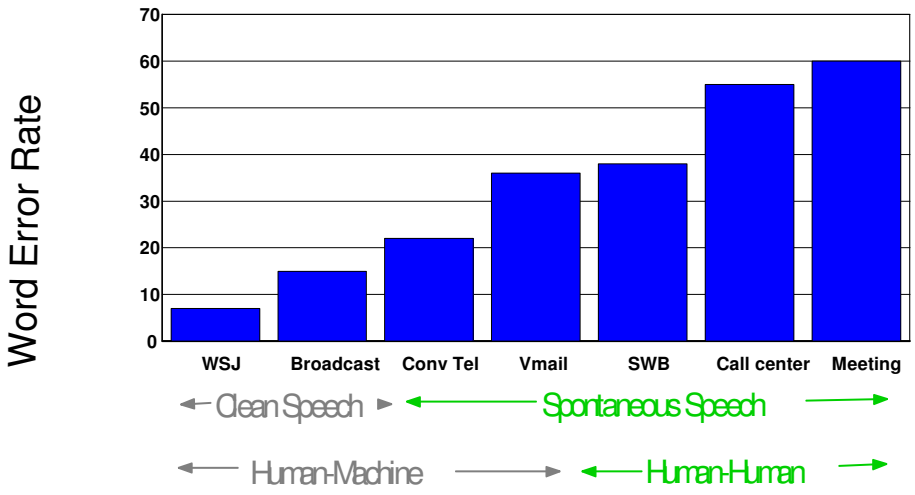
Human Language Technologies:
voice enabled mobile transaction and interaction services in a global community

- Automatic Speech Recognition
  - Telephony Speech Recognition: Conversational Transactions
  - Embedded Speech Recognition: Pervasive Computing - handheld & mobile devices
  - Superhuman Speech Recognition: Large Vocabulary
  - Audio-Visual Speech Recognition: Noise Robust
- Natural Language Understanding & Free Form Dialog
  - Conversational Interaction
- Expressive Text-To-Speech: Human-Sounding TTS
- Conversational Biometrics
  - Security: Voice Identification and Verification Agent
- Speech-to-Speech translation
  - Multilingual conversations, transactions, information access

Major Conversational Technologies

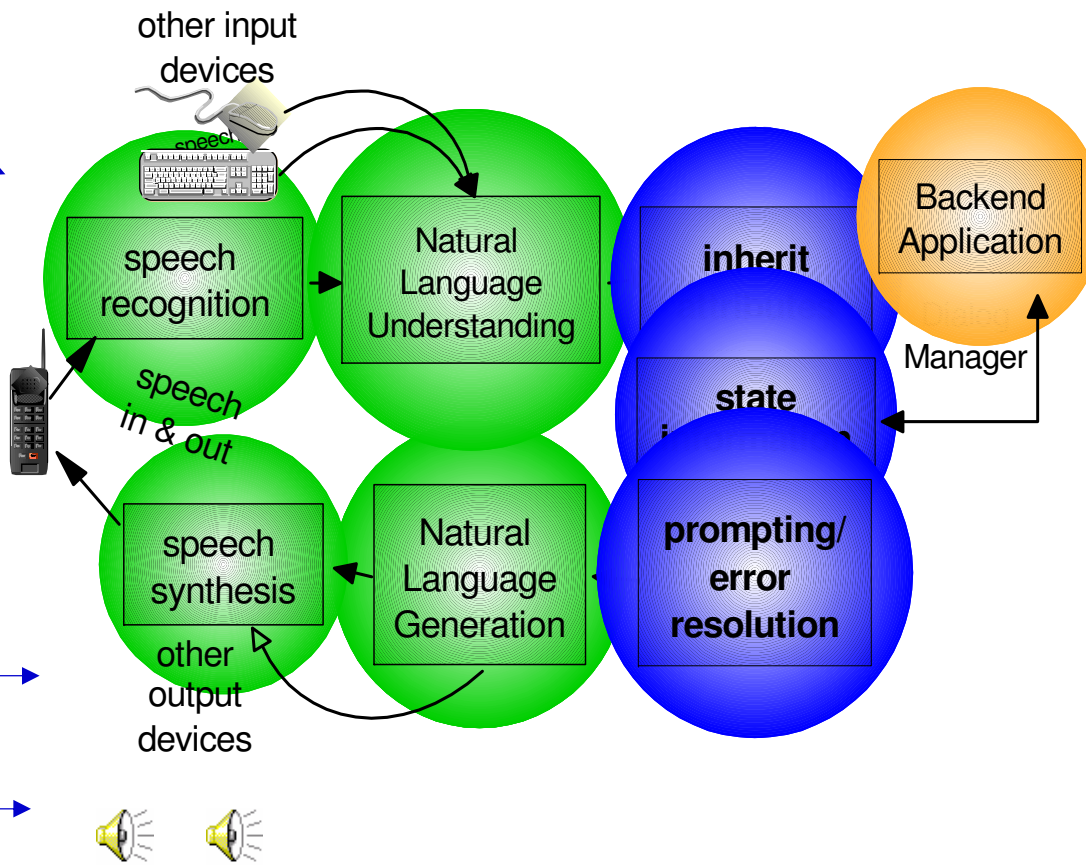
- Automatic Speech Recognition
  - Telephony Speech Recognition: Conversational Transactions
  - Embedded Speech Recognition: Pervasive Computing: handheld & mobile devices
  - Superhuman Speech Recognition
    - Transparent to user, no feedback, across channel, domain & environment
  - Audio-Visual Speech Recognition: Noise Robust
- Natural Language Understanding & Free Form Dialog
- Expressive Text-To-Speech
- Conversational Biometrics
  - Security: Voice Identification and Verification Agent
- Speech-to-Speech translation
  - Multilingual conversations, transactions, information access





Major Conversational Technologies

- Automatic Speech Recognition
  - Telephony Speech Recognition: Conversational Transactions
  - Embedded Speech Recognition: Pervasive Computing: handheld & mobile devices
  - Superhuman Speech Recognition
  - Audio-Visual Speech Recognition: Noise Robust
- Natural Language Understanding & Free Form Dialog
  - Conversational interaction
- Expressive Text-To-Speech
  - Human sounding TTS
- Conversational Biometrics
  - Security: Voice Identification and Verification Agent
- Speech-to-Speech translation
  - Multilingual conversations, transactions, information access



Major Conversational Technologies

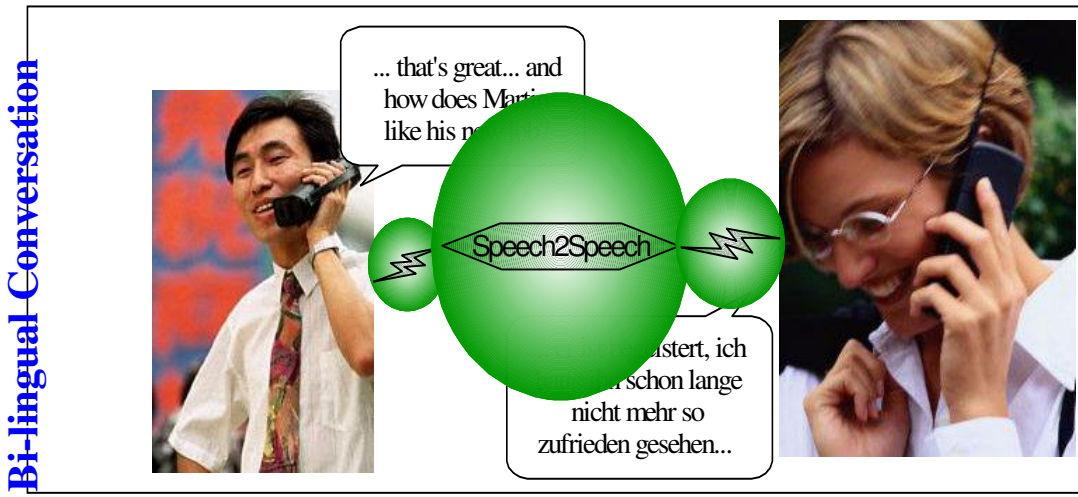
- Automatic Speech Recognition
  - Telephony Speech Recognition: Conversational Transactions
  - Embedded Speech Recognition: Pervasive Computing: handheld & mobile devices
  - Superhuman Speech Recognition
  - Audio-Visual Speech Recognition: Noise Robust
- Natural Language Understanding & Free Form Dialog
  - Conversational interaction
- Expressive Text-To-Speech
  - Human sounding TTS
- Conversational Biometrics
  - Security: Voice Identification and Verification Agent
- Speech-to-Speech translation
  - Multilingual conversations, transactions, information access

Major Conversational Technologies

- Automatic Speech Recognition
  - Telephony Speech Recognition: Conversational Transactions
  - Embedded Speech Recognition: Pervasive Computing: handheld & mobile devices
  - Superhuman Speech Recognition
  - Audio-Visual Speech Recognition: Noise Robust
- Natural Language Understanding & Free Form Dialog
  - Conversational interaction
- Expressive Text-To-Speech
  - Human sounding TTS
- Conversational Biometrics
  - Security: Voice Identification and Verification Agent
- Speech-to-Speech translation
  - Multilingual conversations, transactions, information access

Major Conversational Technologies

- Automatic Speech Recognition
  - Telephony Speech Recognition: Conversational Transactions
  - Embedded Speech Recognition: Pervasive Computing: handheld & mobile devices
  - Superhuman Speech Recognition
  - Audio-Visual Speech Recognition
- Natural Language Understanding & Free Form Dialog
  - Conversational interaction
- Expressive Text-To-Speech
  - Human sounding TTS
- Conversational Biometrics
  - Security: Voice Identification and Verification Agent
- Speech-to-Speech translation
  - Multilingual conversations, transactions, information access



Customer Problems

- Different languages spoken by people living in different regions or even by different ethnic groups living in the same region

- Language barriers cause…
  - Business losses for companies & inconvenience for international business or tourism travelers
    - Difficulties for travelers include transportation, accommodation, shopping, banking, etc.
    - Need for clear communication in international corporate meetings and conference calls
    - Requirement for Call Center human & machine-based self-services for multi-lingual speakers & travelers
  - Life threatening issues for individuals or large groups of people
    - Humanitarian personnel providing help to people in under-developed countries or emerging crisis areas
    - Foreign patients (travelers or immigrants) seeking emergency medical help
    - Law enforcement and security personnel talking to foreign travelers at airports, coastal and land checkpoints for security and anti-terrorism purposes

Technical Challenges

- Translation of speech (as opposed to written text) is greatly complicated
  - Spontaneously spoken speech often is ill-formed, includes non-grammatical disfluencies
  - Text obtained by the speech recognition decoder includes recognition errors caused by imperfect speech recognizer and background noises
- Current ASR (speech recognition)
  - not robust to narrow bandwidth, noise, spontaneous, conversational speech
- Current MT (machine translation): not designed to handle output of ASR system
  - Recognition errors
  - Spoken language: different from written language
    - Non-grammatical disfluencies
    - Imperfect syntax
    - Lack of formal characteristics of text: no punctuation or paragraphing
  - Translated text must be "speakable" for oral communication
    - not adequate to just translate keywords
- ASR-MT interactions are inevitable
  - much more complex than building individual components and gluing them together
  - has to be addressed specifically

DARPA CAST (Babylon) Program

- Main goals:
  - Enhance situational awareness of warfighters in different environments by enabling them to converse in multiple languages
  - Build functional prototypes ready for limited production
- Steps:
  - Phraselator: uni-directional, constrained to fixed phrases
  - Bi-directional, constrained
  - Bi-directional, free form
    - for force protection and medical domains
    - Participants: IBM, SRI, HRL/USC, CMU
- IBM's role:
  - English-Mandarin
  - Produce laptop & PDA prototypes
  - Explore new approaches for S2S

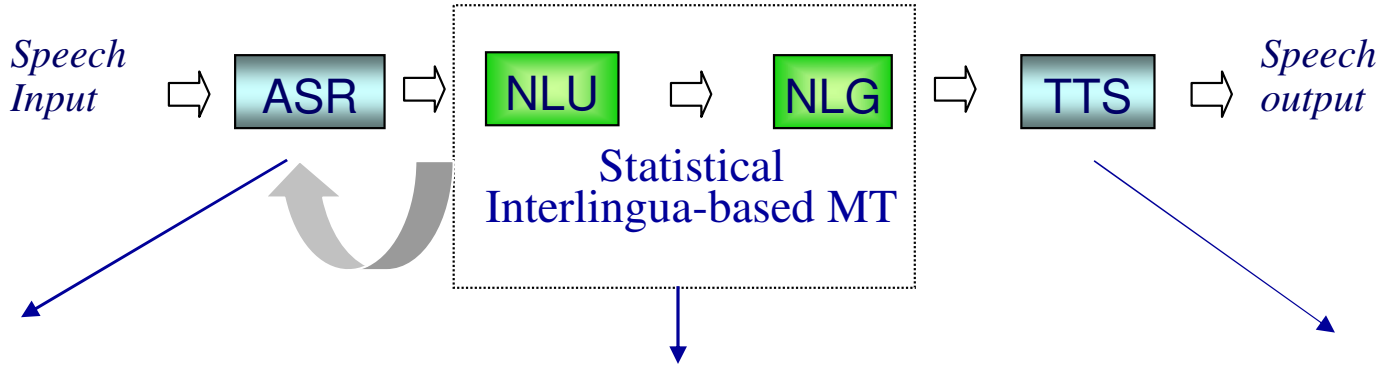
DARPA Live Test – Feb 2004

- DARPA Live Test
  - S2S system mediated bilingual conversational - still spontaneous speech
    - Speakers were more cautious
    - Speakers were requested to speak one sentence at a time
    - Speakers got feedback from system, can repeat or adjust if see errors
- IBM MASTOR system
  - ASR Performance
    - English: WER: 8.86%
    - Chinese: CER: 9.48%
  - Speech-to-Text Translation Performance
    - Accuracy: above 85% communication rate

Technical Approaches

- Strategy:
  - Build upon speech and language technologies
  - Evolve from limited domains (medical assistance, tourism & phone banking) to broader domains (international meetings)
  - Focus on meaning preservation, rather than exact translation
  - Measure progress in multiple fashions
    - Explicit recognition and translation accuracy metrics
    - More importantly, success of communication of ideas, thoughts, and concepts between humans.
- Approaches
  - Meaning preserving and understanding based translation
    - Concept & emotion analysis & translation
  - Coupling ASR & NLU - semantic ASR
  - Unified modeling for recognition, understanding & translation - long-term

Advanced Speech-to-Speech Translation Technique

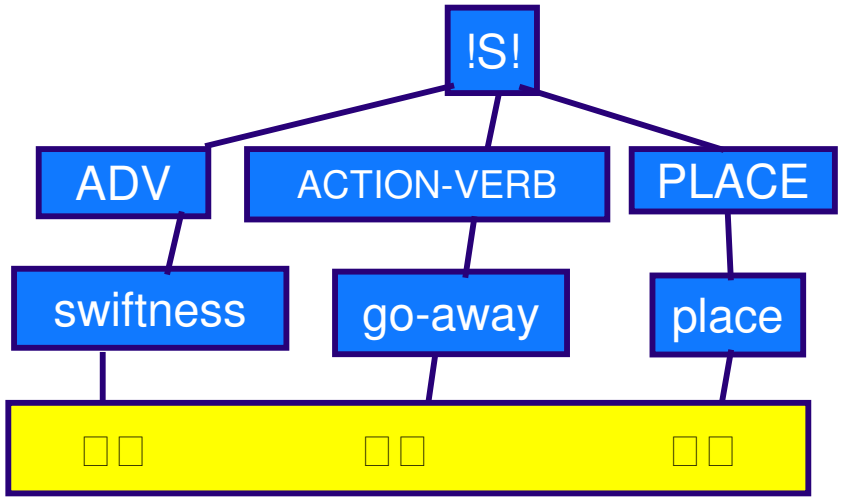


- The speech input is decoded by a large-vocabulary automatic speech recognizer (ASR) into written words.
  - The output of the speech recognizer can be analyzed to determine if the content is within the domain of interest.
- The spoken sentences are translated into the target language.
  - The sentence is parsed by a natural language understanding (NLU) engine.
  - Perform concept and word translation and perform natural language generation (NLG).
- If desired, the text can then be rendered into speech using a TTS (Text-to-Speech) engine.

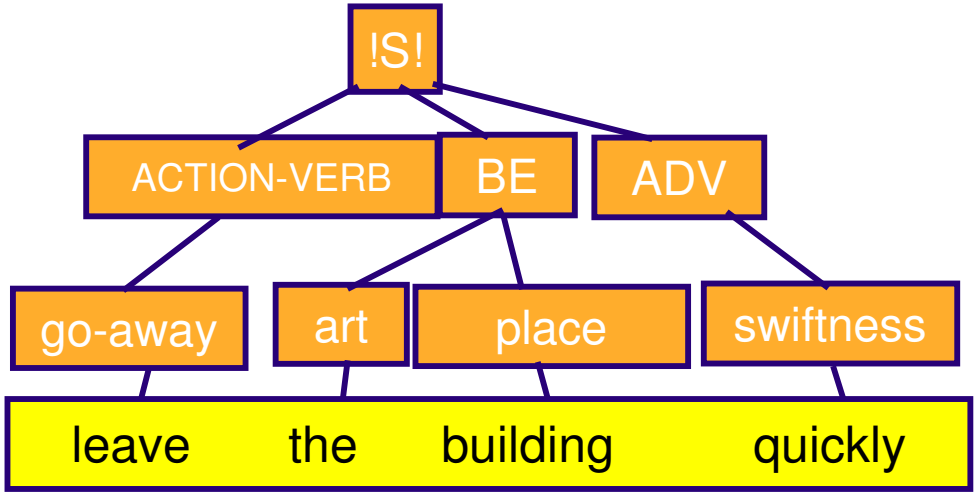
Concept based translation

- statistical NLU & NLG models

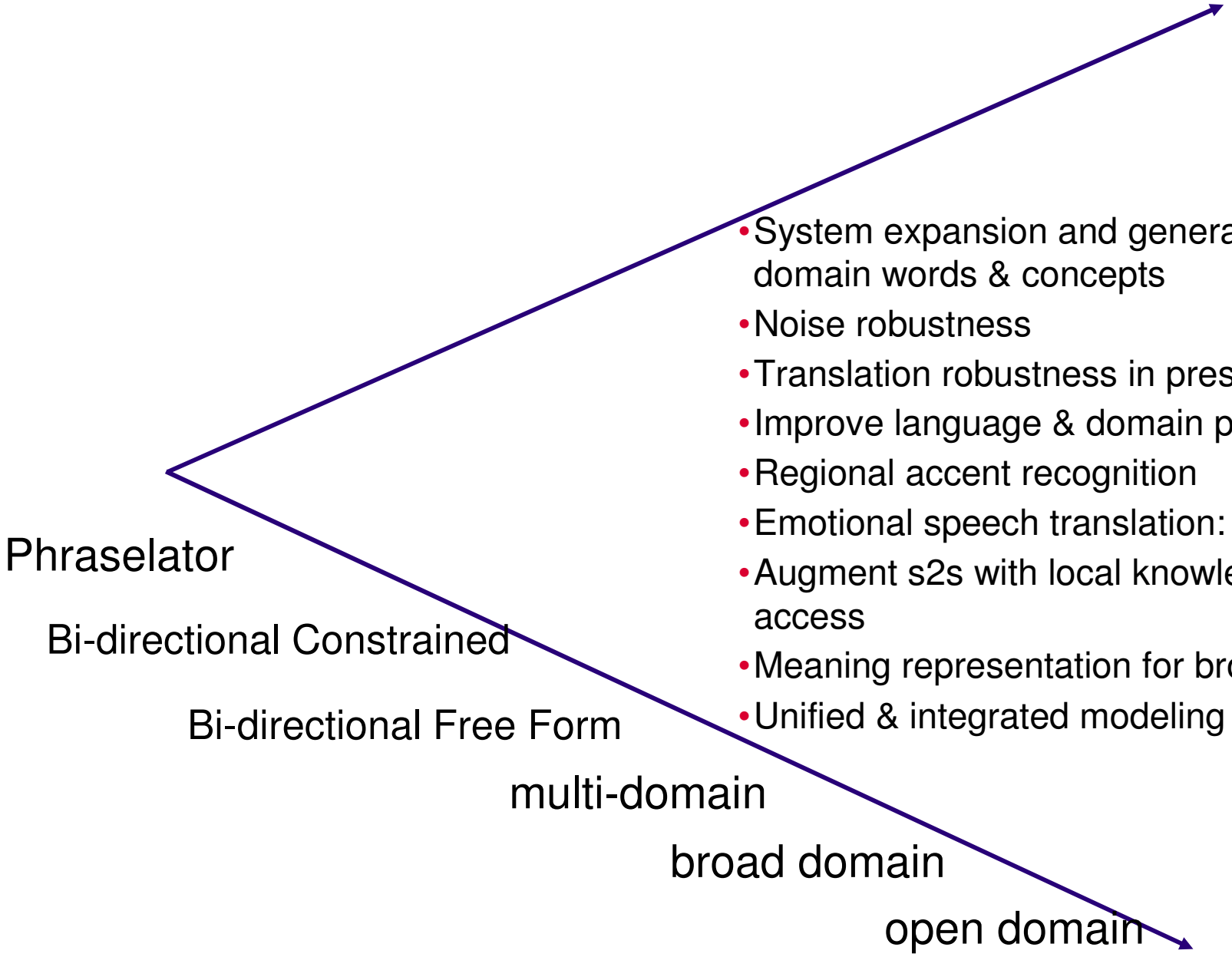
NLU: analysis
input sentence:



NLG: generate
output sentence:



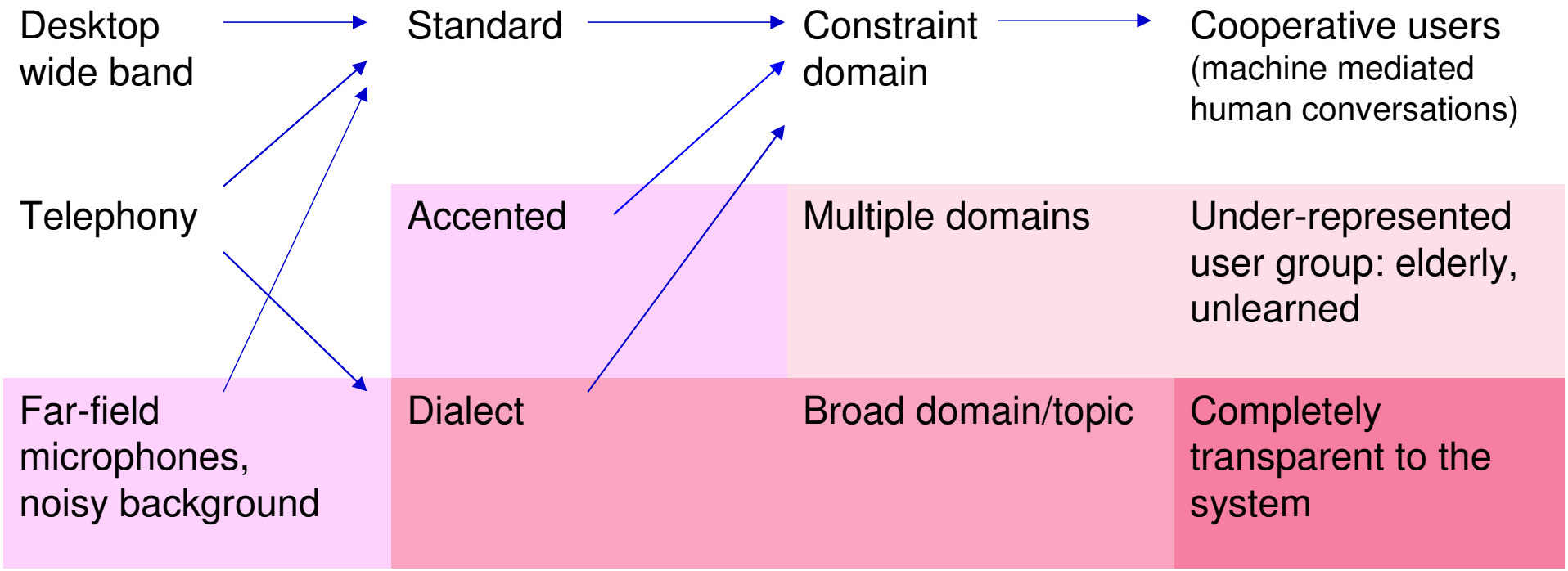
Short-term & long-term challenges



- System expansion and generalization for out-of-domain words & concepts
- Noise robustness
- Translation robustness in presence of ASR errors
- Improve language & domain portability
- Regional accent recognition
- Emotional speech translation: ASR & TTS
- Augment s2s with local knowledge, information access
- Meaning representation for broad & open domains
- Unified & integrated modeling for ASR, NLU & MT

Challenges for Speech Translation

Acoustic conditions	Dialect/accent	Domains	Speaker awareness
Desktop wide band	Standard	Constraint domain	Cooperative users (machine mediated human conversations)
Telephony	Accented	Multiple domains	Under-represented user group: elderly, unlearned
Far-field microphones, noisy background	Dialect	Broad domain/topic	Completely transparent to the system



Laptop & Desktop 2-way MASTOR System Interface

IBM MASTOR System Version 2.1

IBM
Multilingual Automatic
Speech-to-Speech Translator

Click to start >>>

Translation | Configuration

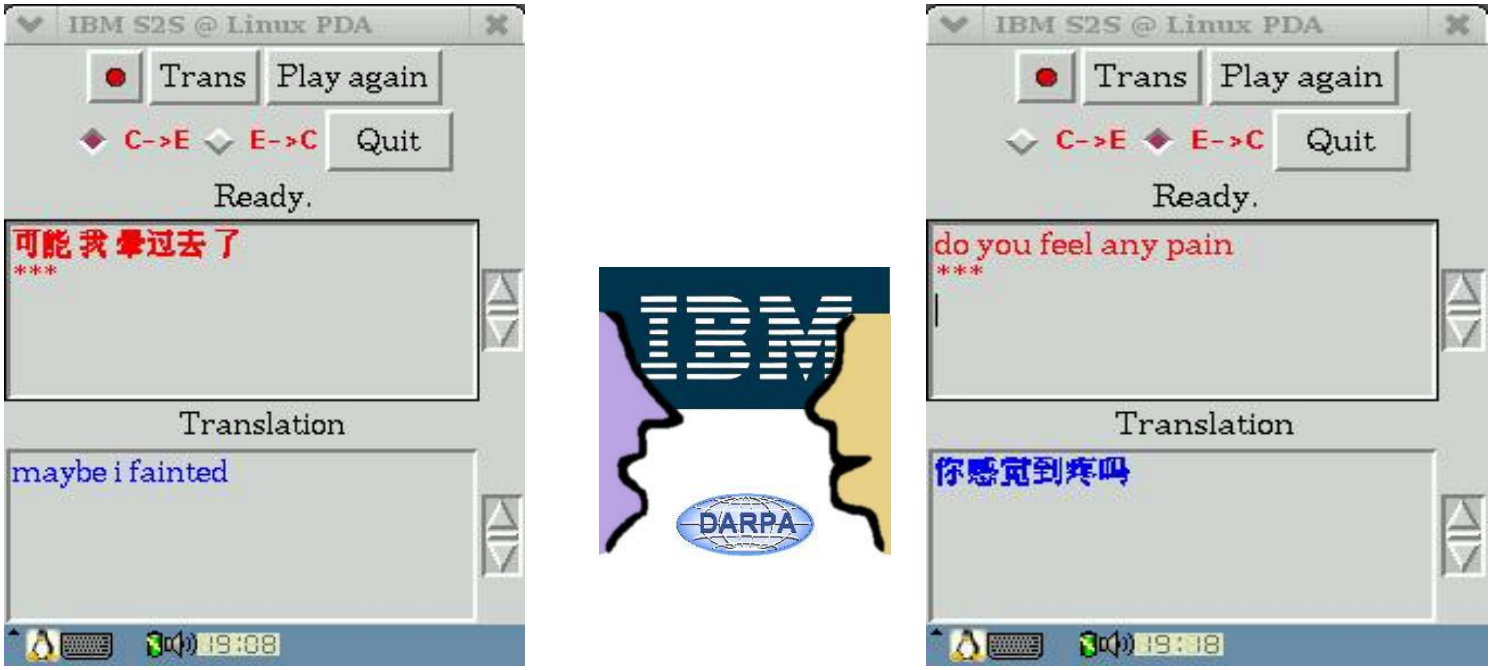
LANGUAGES | English-to-Chinese | Simulate Response | File

Speech I/O Not Activated

English	put your cellphone in this container
Chinese	放 你的 手机 入 这 容器
Back Translation	(place/put/move) (your) (cellphone)(into) (the) (container)

System Ready | Translated | Copyright (c) 2004 IBM Corp.

Handheld 2-way MASTOR System Interface



- PDA: 200MHz CPU, 64MByte memory
- Seamless switch between two directions of translations
  - Two-way engines, shared memory
  - English to Chinese (E->C) vs. Chinese to English (C->E)
- Menu bar control or Push to talk button to control system